Next-generation biology

da Fonseca, Rute R.; Albrechtsen, Anders; Themudo, Gonçalo Espregueira; Ramos-Madrigal, Jazmín; Sibbesen, Jonas Andreas; Maretty, Lasse; Zepeda-Mendoza, M. Lisandra; Campos, Paula F.; Heller, Rasmus; Pereira, Ricardo J.

DOI:
10.1016/j.margen.2016.04.012

Marine Genomics 30 (2016) 3–13

Contents lists available at ScienceDirect

Marine Genomics

journal homepage: www.elsevier.com/locate/margen

Next-generation biology: Sequencing and data analysis approaches for non-model organisms

Rute R. da Fonseca [a,*], Anders Albrechtsen [a], Gonçalo Espregueira Themudo [c], Jazmín Ramos-Madrigal [b], Jonas Andreas Sibbesen [a], Lasse Maretty [a], M. Lisandra Zepeda-Mendoza [b], Paula F. Campos [b,d], Rasmus Heller [a], Ricardo J. Pereira [b]

[a] The Bioinformatics Centre, Department of Biology, University of Copenhagen, Copenhagen, Denmark
[b] Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark
[c] Section of Forensic Genetics, Department of Forensic Medicine, University of Copenhagen, Copenhagen, Denmark
[d] CIMAR/CIIMAR, Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Rua dos Bragas 177, 4050-123 Porto, Portugal

ARTICLE INFO

Article history:
Received 30 November 2015
Received in revised form 23 March 2016
Accepted 26 April 2016
Available online 13 May 2016

Keywords:
RADseq
RNAseq
Targeted sequencing
Genotype likelihoods
Comparative genomics
Population genomics

ABSTRACT

As sequencing technologies become more affordable, it is now realistic to propose studying the evolutionary history of virtually any organism on a genomic scale. However, when dealing with non-model organisms it is not always easy to choose the best approach given a specific biological question, a limited budget, and challenging sample material. Furthermore, although recent advances in technology offer unprecedented opportunities for research in non-model organisms, they also demand unprecedented awareness from the researcher regarding the assumptions and limitations of each method.

In this review we present an overview of the current sequencing technologies and the methods used in typical high-throughput data analysis pipelines. Subsequently, we contextualize high-throughput DNA sequencing technologies within their applications in non-model organism biology. We include tips regarding managing unconventional sample material, comparative and population genetic approaches that do not require fully assembled genomes, and advice on how to deal with low depth sequencing data.

1. Introduction

High-throughput sequencing, more broadly referred to as next-generation sequencing (NGS), has become essential for modern day research within biological sciences, particularly in evolutionary biology. Since the assembly of the first complete genome using Sanger capillary sequencing in 1977 (Sanger et al., 1977), large technological improvements have been made and different methods have been developed. These methods aim to increase the sequencing throughput as well as the quality and length of the reads, while decreasing the time and cost of the process in such a way that it seems that virtually any biological question can be asked given enough data. However, the increase in data production comes with a cost: the corresponding data analysis approaches require a more detailed knowledge on the caveats and drawbacks of each method.

The literature of evolutionary biology has traditionally been dominated by model species such as mammals and drosophilids, for which fully sequenced and well-annotated genomes have been available for years. The recent advent of high-throughput sequencing opened the use of genomic approaches to the study of non-model organisms, allowing the test of generalizations based on a limited number of model species, and unlocking new research programs in fields related to evolutionary biology, such as phylogenomics and population genomics. The choice of the sequencing approach needs to take into account the evolutionary time scale of the biological question (Fig. 1). For example, transcriptome data (RNA-seq) has been used to produce hundreds of protein alignments that resolved deep phylogenetic relationships in Metazoans (Smith et al., 2011) and in plants (Wickett et al., 2014), providing key insights into how characters such as development, morphology or genome structures evolve throughout the tree of life (Dunn et al., 2014). For taxa that have diverged at relatively deep time scales, up to hundreds of millions of years of evolution, targeted sequencing of highly conserved genomic regions (named ultra-conserved elements or UCEs) have been used to establish well-resolved phylogenies of large species radiations such as Amniotes (Faircloth et al., 2012), vertebrates (Lemmon et al., 2012), birds (Prum et al., 2015) or mammals (McCormack et al., 2012). For bacterial species, which have a simple genomic structure, whole genome sequencing has been used to ask similar questions (Ziemert et al., 2014). Targeted sequencing can be extended to micro-evolutionary time scales by designing targets specific for coding and non-coding genomic regions, based on partial genomes of

* Corresponding author.
E-mail address: fonseca@binf.ku.dk (R.R. da Fonseca).

http://dx.doi.org/10.1016/j.margen.2016.04.012

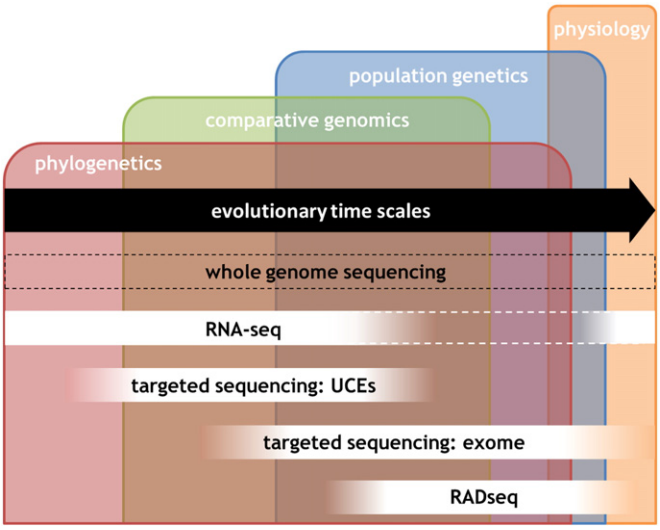

Fig. 1. Application of different high-throughput sequencing methods to different evolutionary time scales. Research applications related to evolutionary biology (colored polygons) address biological questions at different, but overlapping, evolutionary time scales (black arrow), spanning from hundreds of millions of years of evolution between Phyla (left), to generations between populations, individuals or cells (right). In the absence of whole genome sequencing, different high-throughput sequencing methods (white bars) provide a cost- and time-efficient alternative for non-model species. Benefits and limitations of each method depend on the time scale relevant to each biological question.

closely related species (e.g. the domestication of maize (da Fonseca et al., 2015), and diversification of palm trees (Heyduk et al., 2016)). At shorter evolutionary time scales, methods such as Restriction-site Associated DNA sequencing (RADseq), in which thousands of genomic regions spread throughout the genome are sequenced, have been used to establish phylogenetic relationships between closely related species and populations, despite the large confounding effect of interspecific gene flow and incomplete lineage sorting (e.g. diversification and hybridization in oaks (Eaton et al., 2015)). In order to generate summary statistics for population genetics in the absence of a reference genome, Gayral et al. (Gayral et al., 2013) established a pipeline for transcriptome data that controls for paralogue genes and for variation in gene expression among individuals and loci. Using this method, population genomics studies across animals have shown that levels of genetic diversity within a species seem to be largely determined by its ecological strategy, such as propagule size and fecundity (Romiguier et al., 2014), rather than geographic range or invasive status. The same methods have been applied to data from multiple populations within the same species, to understand how genetic drift and positive selection contribute to divergence patterns across the genome (Tsagkogeorga et al., 2012), and to identify genes repeatedly evolving under selection across multiple independent populations (Pereira et al., n.d.). Analysis of polymorphisms (e.g. F_{ST}) estimated from transcriptome data from different species (Renaut et al., 2013) or from populations within a species (Carneiro et al., 2014) have also been used to determine genomic areas of high differentiation that could harbor genes involved in local adaptation and genetic barriers to gene flow, offering insights into the genetic basis of speciation. At a more recent evolutionary time scale, RNA-seq can be used to describe genes involved in physiological adaptation of populations to different environments (e.g. in corals (Barshis et al., 2013)) or to physiological variation between individuals or cells.

Despite these broad applications of high-throughput sequencing, choosing the most appropriate method to address a specific biological question requires considering the benefits and limitations of each method.

2. Sample quality and preparation

The type, quality and quantity of the tissue samples used in genomic analyses has a great impact in the final results, both in terms of quantity and quality of reads obtained after sequencing. The success of high-throughput sequencing approaches is highly dependent on the use of high molecular weight DNA/RNA samples, and this is only possible if fresh or carefully stored tissue samples are used for nucleic acid isolation. Frozen collection and transportation can be challenging, and in fields like marine genomics, this is not always feasible. For some species (mammals, birds, fish) tissues like blood or muscle might be a good source of DNA. Plant tissues rich in resins, gums, and polyphenolics should be avoided (Abu Almakarem et al., 2012). When the specimen size is very small (e.g. small marine invertebrates) the whole specimen or even a pool of several specimens might be required to obtain enough genetic material for subsequent analyses. For microbes (bacteria, fungi, diatoms, microalgae) single species isolates are needed to ensure the required amounts of pure DNA for NGS sequencing (but laboratorial culturing is only possible for a restricted number of species).

Ideally, when in the field, samples should be collected and immediately stored either in liquid nitrogen (preferred), at −20 °C (using a freezer or dry ice) or in a chemical preservative (RNAlater type solution), a solution that rapidly permeates the tissue and protects cellular nucleic acids in unfrozen tissue samples, as these materials are susceptible to fast post-mortem degradation or degradation following collection from living specimens. If possible, several subsamples should be obtained per specimen as a back-up procedure in case something goes wrong between collection and sample processing, or to allow for high coverage sequencing. In cases where the species of interest is extinct, very rare or difficult to collect in the field, museum specimens might also be used to obtain genetic material. Such specimens are also valuable for time series analysis (e.g. (Bi et al., 2013)). Tissue samples stored in ethanol (internal organs, muscle), dried (beaks, bones), taxidermized (skin, nails, hair), or frozen in tissue banks are alternatives to freshly collected samples. The DNA obtained from these samples is usually of inferior quality (lower molecular weight and concentration) and the amount of external contaminants will be higher.

Crucial for obtaining high molecular weight DNA is also the choice of extraction method used for nucleic acid isolation, appropriate for tissue type and preservation method (Campos et al., 2009). Another extremely important preliminary step is to obtain all legal permits and documentation associated with collection of samples. Large sequencing consortiums like Genome 10 K have established very strict protocols for tissue collection and storage (Wong et al., 2012) to maximize the information obtained for each specimen.

3. Restriction-site Associated DNA sequencing (RADseq)

RADseq (Restriction-site Associated DNA sequencing) was developed as a method for the simultaneous discovery and genotyping of tens of thousands of genome-wide markers through a reduced representation protocol (Baird et al., 2008). It can be used for population genetic analyses or for building genetic maps. Two features of RADseq have made it popular for population genetic studies on non-model species: i) it does not require a reference genome and ii) it provides a cost-efficient way of genotyping genome-wide markers in many individuals. RADseq is flexible in that the restriction enzyme can be chosen so as to fit the desired reduction factor, i.e. the proportion of the genome that is sequenced. Hence the whole continuum between sequencing a few loci (e.g. a few tens of thousands) at high coverage or many loci (e.g. many hundreds of thousands) at lower coverage is accessible in a RADseq study. For example, a single Illumina lane can be used to sequence 100 individuals at 150,000 RAD loci, providing up to 10× mean coverage per locus per sample. Such data has been used to identify fresh-water adaptation in sticklebacks (Hohenlohe et al., 2010) and to infer the phylogeny of African cichlids (Wagner et al., 2013). While

RADseq is an attractive method for many applications, population genetic analyses of RADseq data are challenged by several complications. For example, it has been shown that restriction site length polymorphism and PCR duplication can introduce bias in locus sequencing depth (Davey et al., 2013; Gautier et al., 2013). Furthermore mutations in the enzyme recognition sites can lead to allelic dropout, which can lead to underestimation of heterozygosity and other analytical artifacts. Variations of the original RADseq protocol have been proposed to alleviate some of these concerns (ddRAD (Peterson et al., 2012), ezRAD (Toonen et al., 2013), 2b-RAD (Wang et al., 2012)), and the list of studies developing improvements to RADseq pipelines is growing. Overall, the consensus is that the challenges can be overcome by prudent filtering (Davey et al., 2013) and that RADseq is a useful technique for species with little or no genomic resources. However, it still has some severe limitations deriving from the lack of positional information for each RAD locus, like other reference-free methods. This can limit the researches possibility to adjust for linkage disequilibrium (LD) for commonly used measures used in population genetics, such as F_{ST}, nucleotide diversity, D-statistics, estimated admixture proportions and many others. Additionally, it will be hard to detect signs of selection since selection scans or region-wise analysis is not possible. A comprehensive study of the peculiarities of RADseq data – including biases related to genotype calling – compared with regular shotgun sequencing is still lacking and would be highly useful.

4. Targeted sequencing

Until recently, the development of genetic markers that enable macroevolutionary and population genetic studies was a significant hurdle when dealing with non-model organisms. In particular, the lack of extensive genomic resources meant that significant time and resources were placed in building and sequencing EST or BAC libraries, or in either trial-and-error primer design (as in the development of markers for phylogenetic studies, e.g. (Espregueira Themudo et al., 2009)), or marker testing and selection. Targeted sequence capture facilitates marker development for a single species. In fact, as probe specificity does not need to be high, the same set of probes can be used for multiple related species (as was done in the plant genus *Inga* (Nicholls et al., 2015)). Depending on the evolutionary time scale under study, capture probes can be based on genic regions from a reference genome of the same or a closely related species or derived from UCEs. It is important to consider how conserved the target regions will be in the species under study. High conservation will lead to higher capture efficiency but can potentially result in limiting variation for the downstream evolutionary analysis.

Several types of technologies allow for targeted sequence capture, and can be classified according to the enrichment method: hybridization-based, PCR-based or molecular inversion probe-based (Mamanova et al., 2010). Each of these has their own advantages and disadvantages, and several commercial products using any of these methods (or derivations) are available, such as Agilent's SureSelect or Haloplex, MYcroarray's MYbaits or Roche NimbleGen's SeqCap. For studies that aim targeting specific candidate loci, there are several protocols available for do-it-yourself capture (see (Grover et al., 2012)) for a list of references).

As capture reduces the genomic sequence space compared to whole genome sequencing (WGS), and it enables the multiplexing of several individuals, contributing to further reducing the overall costs of sequencing per sample (Grover et al., 2012). Furthermore, it reduces the complexity of the analysis compared to WGS as only the required number of genes is sequenced. Reduction of sequencing costs per sample allows the inclusion of more taxa in any given study. In phylogenetic studies this is particularly important, as inadequate taxon sampling introduces artifacts such as long branch attraction (e.g. (Prum et al., 2015)).

By allowing better spatial (and temporal) sampling, targeted sequencing also enables more detailed reconstruction of dispersal routes and gene flow between varieties or subspecies (Nadeau et al., 2012; da Fonseca et al., 2015). This is essential to better understand the evolution of adaptive phenotypes, as we can now pinpoint when and where they developed, and how they spread.

5. Transcriptome sequencing (RNA-seq)

RNA-seq can be used both with and without a reference genome, and has different applications along the evolutionary time scale (Fig. 1). As it is composed of mostly coding regions, it can be used for reconstruction of deep-phylogenies, where often only a few transcripts with high-confidence orthologs are used. Yet, like other reduced genome representation methods mentioned above, it does not contain information about linkage among genes. Also, because coverage is dependent on gene expression, the uncertainty of genotype calling strongly varies across genes and should be taken into account (see (Gayral et al., 2013)). In addition, the breadth and amount of individual transcripts provide insight into the regulation of biochemical processes and pathways, and one can use RNA-seq to assess differential gene expression in different tissues of the same individual, between individuals with different phenotypes or under different environmental stresses. Using RNA-seq data to infer differential gene expression in non-model organisms is more challenging than to infer gene sequences. Biological variance is typically much higher in field studies compared to studies based on inbred organisms or cell lines, demanding much higher samples sizes in order to achieve similar statistical power (Todd et al., 2016).

If a reference genome is available, it is possible to both call variants (e.g. (Piskol et al., 2013)) and identify differentially expressed genes (e.g. (Love et al., 2014); requires gene annotation). This can be done directly from alignments obtained using dedicated RNA-seq tools such as TopHat2 (Kim et al., 2013), HiSat (Kim et al., 2015) or STAR (Dobin et al., 2013) that allow for spliced reads. It is important to note that the default behavior of splice-aware aligners is to favor splice junctions that match established human splice-site motifs. Consequently, it may be necessary to adjust the splice-site scoring parameters of the aligner if the organism under study uses splice sites that deviate from the human motifs. STAR (Dobin et al., 2013) and HiSat (Kim et al., 2015) include options on how to score canonical versus non-canonical splice sites. However, if a reference genome is not available or resolution of unannotated transcripts such as alternative splice-variants is needed, full-length transcripts must first be reconstructed from the RNA-seq data. This is generally done in two steps. First, the overall gene structure is extracted from the read sequences and represented as “splice-graphs”, where nodes and vertices represent exons and exon-exon junctions, respectively. By definition, the transcripts that actually generated the data correspond to paths in these graphs. The second step is the identification of the correct transcripts among all possible paths, which is non-trivial when multiple transcripts share longer stretches of sequence e.g. due to alternative splicing or in the presence of close paralogs or contaminant orthologs.

The approach used for constructing the splice-graphs distinguishes two classes of assembly algorithms. Reference-based approaches work by first aligning the RNA-seq reads to a reference genome and then building the graph by combining reads that overlap on the reference genome. De novo assembly algorithms build the splice-graph directly by comparing read sequences and thus do not make use of a reference genome. By leveraging the reference sequence to bridge regions with low coverage, reference-based approaches, programs such as Cufflinks, as well as more recent approaches like those in Bayesembler (Maretty et al., 2014) and StringTie (Pertea et al., 2015), will generally be able to assemble more lowly expressed transcripts than de novo approaches (Trapnell et al., 2010; Maretty et al., 2014; Pertea et al., 2015). But reference-based methods generally ignore variation from the reference sequence observed in the reads (e.g. SNVs) as they are focused on

determining only the overall exon-structure of transcripts. Hence, algorithms based on de novo graph construction such as Trinity, Oases, and the more recent SOAPdenovo-Trans and Bridger, are thus preferable when such information is required e.g. for population genetic inferences or when no suitable reference sequence is available (Grabherr et al., 2011; Schulz et al., 2012; Xie et al., 2014; Chang et al., 2015).

Most assembly algorithms use graph optimization approaches to find the correct transcripts in the graph. Trinity, Oases, and SOAPdenovo-Trans use more heuristic approaches, whereas Cufflinks, StringTie and Bridger apply more rigorously founded algorithms. The latter three methods tend to produce simple assemblies as they explicitly try to find sparse solutions, whereas the former tend to produce significantly more transcripts. Finally, the Bayesembler solves the graph inference problem using an alternative, probabilistic approach that allows it to also estimate a confidence score for each transcript. Hence, the larger transcriptome estimates produced by de novo assemblers reflect both differences in graph construction (e.g. due to inclusion of heterozygotic SNVs) and in graph inference. Finally, contaminant transcripts, which may be substantial (Strong et al., 2014), may also increase the relative complexity of de novo assembled transcriptomes as they will likely be eliminated in the reference-based approach. Such contaminants can be removed by e.g. aligning the assembled transcripts to a database of expected contaminants before the results are used for downstream analysis.

Comparisons between the different reference-based and de novo approaches can be found in Hayer et al. (2015) and Yang and Smith (2013); Li et al. (2014), respectively, although it is important to choose the method that is the most adequate for the question and the available sample set. Even determining which method is the most accurate is non-trivial as good reference sets are lacking, but metrics based on previously annotated transcripts (Maretty et al., 2014), assembly likelihoods (Li et al., 2014) and in vitro transcription (Hayer et al., 2015) have been proposed.

When using de novo assembled transcriptomes, one should also assess the overall quality of the final assembly. The two most relevant metrics are the completeness and the degree of fragmentation of the assembly. Completeness refers to the proportion of assembled transcripts compared to the total set of available transcripts in the cell. This measure is highly dependent on the experimental conditions and it is commonly assessed by determining the presence or absence of transcripts originated from well-established housekeeping genes (O'Neil and Emrich, 2013). The contiguity of the transcriptome assembly, which is expected to consist of gene-sized contigs, can be estimated using e.g. DETONATE (Li et al., 2014) and TransRate (Smith-Unna et al., 2015).

The functional annotation of de novo assembled transcriptomes can be done with same approaches used for annotating genomic gene sets (see below). There are also transcriptome-specific tools, such as Annocript which besides using BLAST to build functional annotations using information from several databases, also calls putative long non-coding RNAs (Musacchia et al., 2015).

6. Sequencing platforms

Due to method-specific features, one can choose the technology that is better suited for a given project. Sequencing platforms that produce a large amount of short reads should be preferred for resequencing organisms for which a reference sequence is already available (e.g. (Chia et al., 2012; Bertolini et al., 2015)). On the other hand, platforms that generate longer reads may be used in combination with the former for de novo sequencing projects (e.g. (Denoeud et al., 2014; Berlin et al., 2015)) or to resolve structural variants (e.g. (Wang et al., 2015; Ummat and Bashir, 2014)).

Currently, the high-throughput of Illumina and the availability of software to analyze its data makes it ideal for projects aimed at resequencing large sample sets. However, sequencing technology research is also moving towards the production of single molecule long reads. For example, the recently released nanopore technology (Cherf et al., 2012) and its scalable MinION platform (Hargreaves and Mulley, 2015) have made it possible to sequence single molecules with no length limit, and to produce data in real-time. However, high error rates in long reads should be decreased, and software intended for corresponding analyses should be refined before such technologies can compete for the leadership. Nevertheless, a combination of different sequencing technologies is sometimes the best strategy to follow (e.g. (Madoui et al., 2015)). Table 1 summarizes the characteristics of current sequencing technologies.

7. De novo genome assembly

Certain projects still require a genome assembly, which very often is no more than a low quality draft.

Early genome projects aimed at a very high accuracy and contiguity across the genomic assemblies, and a single genome was often resulting from the work of large groups of scientists and many years of data curation. Biological limitations such as recent whole genome duplications and segmental duplications complicate the process of assembly, and time and money are often limiting factors. In particular, the distinction between haplotypes and paralogs resulting from recent and sometimes partial duplication events can be a major problem in species with high heterozygosity (e.g. sea urchin (Sodergren et al., 2006) and oyster (Zhang et al., 2012). This can be partially overcome by using long reads lengths and mate pairs with large insert sizes, but ideally the sample used in genome sequencing should come from an individual with low heterozygosity, potentially an inbred created for that single purpose. Also, whenever possible, the sample should originate from one single individual, given that allelic variation can be increased even more by adding the population variation to the individual variation.

A faulty assembly can impair the evaluation of presence/absence variation, synteny, gene size evolution and protein sequence evolution analyses. Before starting a genome assembly, one must consider which genome from the cell is under study: mitochondrial, plastidial, or nuclear. This is relevant as the presence of mitochondrial or plastidial sequence insertions into the nuclear genome (du Buy and Riley, 1967; Samaniego Castruita et al., 2015) can lead to misidentifications of orthologous genes. Genome assemblies can be reference-based, de novo and hybrid, with the de novo assembly of nuclear genomes posing the greatest laboratory and computational challenges. De novo assembly consists of connecting the high-throughput sequencing reads using algorithms based on mathematical concepts such as de Bruijn graphs (Compeau et al., 2011). Paired-end reads are used to obtain a long contiguous sequence with the short reads, since they allow the program to connect reads very distant from each other (Collins and Weissman, 1984). Thus, high quality de novo assemblies use different insert sizes of paired- (e.g. 170 bp, 500 bp and 800 bp) and mate-end reads (e.g. 2 kbp, 5 kbp, 10 kbp, or even longer) (Gnerre et al., 2011; Geng et al., 2012).

The degree of fragmentation (i.e. how many short unordered and unconnected scaffolds) in the assembled draft genome depends on the species genomic characteristics, such as repetitive regions (e.g. telomeres and centromeres), segmental duplications, GC content that might bias the sequencing protocols, and ploidy. In particular, assembling the genomes of polyploid genomes or species with recent whole genome duplications is a big challenge. The quality of the assembly will have an impact on the type of analyses that can be done. It has been shown, for example, that assemblies with different qualities can produce different gene annotations (Florea et al., 2011). However, depending on the project goals, it might be possible to obtain enough information from a highly fragmented but complete assembly. This is the case if one only needs to identify SNP variation, but only a highly contiguous assembly will be useful for identifying structural variation.

Metrics such as the number and size of the contigs or the N50 (length of a contig such that the sum of the length of all contigs larger

Table 1
High-throughput sequencing technologies.

Sequencing technology/Platform	Detection method	Library types	Maximum read length (bp)	Reds per run (maximum)	Error rate (approximate)	Pros	Cons
454/GS FLX titanium XL+	Pyrophosphate detection (M., 1998)	Single end/ Paired end	1000	1,000,000	0.2% (Shao et al., 2013)	Medium size reads. Errors are well characterized (Shao et al., 2013).	Will be discontinued in 2016. Inaccurate homopolymer detection (M., 1998). Emulsion PCR[a].
454/GS Junior Systems			800	100,000			
Illumina-Solexa/Hiseq 4000	Fluorescence, reversible terminators (Bentley et al., 2008)	Single end/ Paired end	2 × 150	5,000,000,000	0.2–0.8% (Quail et al., 2012)	Widely used. Flexible library preparation methods. High-throughput well suited for resequencing projects. Good characterization of biases (Schirmer et al., 2015).	Short reads. Not optimal for de novo assembly.
Illumina-Solexa/MySeq			2 × 300	25,000,000			
Life Technologies/ SOLiD	Fluorescence di-base probes (McKernan et al., 2009)	Single end/ Paired end	1 × 75/2 × 50	1,400,000,000	0.01% (Buermans and den Dunnen, 1842)	Second most used (van Dijk et al., 2014). Second highest throughput. Each base is read twice, thus decreasing the error rate.	Color space not supported by many mappers. Short reads. Emulsion PCR[a].
Life Technologies/Ion Torrent	Hydrogen ion (pH) sensor (Merriman et al., 2012)	Single end/ Paired end	400	5,500,000	1.8% (Quail et al., 2012)	Short running times.	Bias against AT-rich regions (Quail et al., 2012). Inaccurate homopolymer detection (Buermans and den Dunnen, 1842). Emulsion PCR[a].
PacBio RS II/SMRT	Fluorescence phospho-linked nucleotides (Eid et al., 2009)	Single end	20,000	55,000	13% (Quail et al., 2012)	Longest reads. Good for improving de novo assemblies. Single molecule sequencing.	Low throughput. High cost-throughput ratio. High error rates (Quail et al., 2012).
Oxford Nanorope/ MinION	Electrical sensing (Cherf et al., 2012)	Single end	2000	60,000	30% (Madoui et al., 2015)	Long reads. No GC bias (Cherf et al., 2012). Portable. Scalable. Real-time data. Single molecule. It is possible to read both strands of the DNA sequence.	High error rates. Quality scores only defined by the quality of the alignment to a reference sequence. Existing mapping and assembly software do not deal with long and high error reads.

[a] The emulsion PCR step adds an extra bias. Only 1/3 of the molecules will contain exactly one molecule, and thus be useful during the sequencing process.

or equal to the N50 corresponds to half of the total assembly length) are commonly used to measure the contiguity of an assembly. However, those metrics alone may be misleading in assessing the completeness of the assembly, since incorrectly merged contigs can artificially increase contiguity (Salzberg and Yorke, 2005). In this case, completeness refers to the proportion of the genome included in the assembly and can be directly measured by the total length of the assembly, or indirectly by quantifying the assembled genes, as compared to a set of conserved genes (e.g. (Parra et al., 2007)). Alternatively, previous knowledge about the genome size can be used. Furthermore, discerning real variation from artifacts is a difficult undertaking. REAPR (Hunt et al., 2013) can be used to detect fine-scale inaccuracies, such as artificial substitutions and short indels generated during the assembly process. Consensus quality scores based on sequencing depth and the quality of the individual bases are usually used to assess confidence of each of the assembly positions. General quality metrics for whole genome assemblies can be computed with e.g. QUAST (Gurevich et al., 2013) or GAGE (Salzberg et al., 2012).

Better quality assemblies can be obtained through improvements both in data generation and in the algorithms used to analyze the data (Alkan et al., 2010). For instance, a recent computational and laboratory methodological approach based on a modification of the Hi-C method has been developed that significantly increases the assembly connectivity (Putnam et al., 2015). This method still requires a draft genome as an input, which can be produced by SOAPdenovo (McKernan et al., 2009), MIRA (Buermans and den Dunnen, 1842), Meraculous (van Dijk et al., 2014) and ALLPATHS-LG (Madoui et al., 2015), among others (Baker, 2012; Bradnam et al., 2013). Each has assembly algorithm has its advantages and disadvantages (Bradnam et al., 2013). For example, SOAPdenovo was designed to work with Illumina short reads with low computational memory requirements, but produces highly fragmented assemblies. On the other hand, ALLPATHS-LG produces assemblies with greater contiguity, but it has high computational requirements. Competitions such as the “Assemblathon” (Earl et al., 2011) provide an evaluation of the weaknesses and strengths of a number of de novo assemblers when dealing with same dataset, which in Assemblathon 2 consisted of sequence data from different vertebrate species (a fish, a bird, and a snake) (Earl et al., 2011; Bradnam et al., 2013). Overall, the Assemblathon competitions showed that the assemblies produced by the different programs for the same dataset can be very different, and that some programs work better for some species than for others (Earl et al., 2011; Bradnam et al., 2013).

Following a de novo genome assembly, one can then annotate the regions of the genome that correspond to elements of biological relevance, such as repeat regions, non-coding RNAs or protein-coding genes. Repeat elements can be characterized with de novo identification tools (e.g. RepeatModeler (http://www.repeatmasker.org), LTR_FINDER (Xu and Wang, 2007), or with the use of homology-

based tools that require a reference database, such as RepeatMasker (http://www.repeatmasker.org). When transcriptomes, ESTs or protein databases are available for species that are closely related to that corresponding to the genome assembly, the structure of the genes can be predicted from an alignment of the protein/transcript sequences against the genome by e.g. Scipio (Keller et al., 2008) (that uses BLAT (Kent, 2002)), Exonerate (Slater and Birney, 2005), or GeneSeqer (Brendel et al., 2004). In case there is a large evolutionary distance to the closest annotated species, the use of ExonHunter (Brejová et al., 2009) is advised. Gene prediction can also be done ab initio with programs such as AUGUSTUS (Stanke et al., 2006) and SNAP (Korf, 2004), but these still require some type of homology-based approach for the construction of the training set (Haas et al., 2011). If both ab initio and reference-based predictions are performed, a step for building a non-redundant gene set should be included.

8. Functional annotation

Functional annotation of predicted genes and transcripts can be done by assessing their sequence similarity to proteins of known function, e.g. using approaches based on BLAST (Camacho et al., 2009). Blast2GO (Conesa et al., 2005) is a "biologist-friendly" software that provides a gene ontology (GO) (Ashburner et al., 2000) annotation, and domain-based functions from the InterPro database (Mitchell et al., 2014) (that includes information from several databases such as PROSITE (Sigrist et al., 2010), PRINTS (Attwood et al., 2003), Pfam (Finn et al., 2008), ProDom (Servant et al., 2002) and SMART (Letunic et al., 2015)). InterProScan (Zdobnov and Apweiler, 2001) can be used within Blast2GO or separately. Alternatively, one can use phylogenetic-based approaches, transferring annotations only between one-to-one orthologues and taking advantage of resources such as eggNOG, which has a large set of orthologous groups with functional annotations (Huerta-Cepas et al., 2015).

9. Mapping to a reference genome

Mapping refers to the search for the position in the reference genome that corresponds to each of the reads sequenced. The presence of repetitive regions as well as paralogs complicate that search, and often reduce the number of reads that can be mapped to one single position (uniquely mapped reads).

Table 2 lists some of the currently available programs that are adequate for short-read mapping and their most relevant characteristics.

Mapping algorithms can be broadly classified as hash-based or compression-based, depending on how they store and access the reference genome (Schbath et al., 2012). A common compression-based algorithm applied to mapping programs is the Burrows–Wheeler Transform (Burrows and Wheeler, 1994), which is used to transform (index) the reference genome sequence into a tree-like structure containing all the possible subsequences. Then the first bases of the queries (the reads) are searched on the tree, thus allowing a faster way to access the reference and perform the sequence comparisons. Examples of such software include BWA (Li and Durbin, 2009) and Bowtie (Langmead et al., 2009). Hash-based algorithms identify all the words of a certain length in the reference genome sequence and make a table (hash) with a unique ID for each unique word with the positions of each word in the reference sequence. A similar table can also be produced from the queries and the words from this table are looked up in the reference. These programs usually have large computational memory requirements, and some allow for multi-threading. Examples of such software include MAQ (Li et al., 2008a), SHRiMP (David et al., 2011), and Mosaik (Lee et al., 2014) (Table 2).

Mapping is crucial for the success of the downstream analysis. The quality of the reference will have a great impact in the results. For example, if the assembly does not contain the entire genome, reads corresponding to the missing regions might map to other regions present in the assembly, introducing errors. This can also happen in species with a high content of copy number variants, or differences in copy number variants between individuals. The reference should contain as much unambiguously assembled genomic data as possible, including organelles, as even in targeted-sequencing experiments there is always the presence of genome-wide sequences outside the target regions. Furthermore, if the assembly corresponds to a species that is different from that of the reads, then there could be mapping with high quality scores to paralogous regions, thus it is important to take divergence into account.

10. Analysis for low depth sequencing data

In all NGS projects aiming at population genomics analyses, it is necessary to decide whether one should sequence more individuals at low depth, of fewer individuals at high depth. Often in evolutionary biology, one is interested in comparisons between many individuals, and therefore recent methods developed for low depth data are ideal for population genomic approaches. Low depth is a problem because in NGS data there is often uncertainty or errors in called genotypes. This is also often a problem when using capture methods to enrich certain regions of the genome. Capture methods or even RAD-seq sequencing will often have a very uneven depth distribution such that many of the sites will have low depth.

Calling genotypes for low or even medium depth ($<8\times$) data can cause considerable bias in the downstream analysis. This problem can be illustrated for one of the most useful summary statistics that can be inferred from NGS data, the site frequency spectrum (SFS). Many commonly used statistics in population genetics can be derived from the SFS such as the fraction of variable sites, F_{st} between populations, and neutrality test statistics (e.g. Tajima's D). Based on the SFS it is also possible to infer complex demographic histories as well as past population size changes (Gutenkunst et al., 2009; Liu and Fu, 2015). However, the SFS is especially sensitive to the SNP and genotype calling procedure. The impact of SNP and genotype calling on the site frequency spectrum can be seen in Fig. 2. The plots correspond to results obtained in multiple studies (Nielsen et al., 2011, 2012; Han et al., 2014; Nevado et al., 2014) that are based both on simulations and real data. When calling genotypes it is often useful to first determine if a site is polymorphic based on multiple samples. However, this will bias the SFS because it is easier to call SNPs when the allele frequency is high (Fig. 2E). Therefore, SNP calling will tend to undercall the low frequency categories especially singletons. Choosing not to call SNPs prior to genotype calling will have the opposite effect (Fig. 2C). Due to sequencing errors a fraction of the non-polymorphic sites will be called as singletons and because most of the genome is non-polymorphic this will have a large effect on the SFS. Each genotype will have some quality score, often a probability, associated with the call that indicates the certainty of the genotype. The use of a stringent cutoff on the genotype quality appears to be an appropriate solution, since it increases the accuracy of the genotypes. However, this creates another bias given that it is easier to obtain a large quality for the call of a homozygotic site relative to a heterozygotic site, hence fewer polymorphic sites will be called (Fig. 2F). As shown in Nielsen et al. (2011, 2012); Han et al. (2014); Nevado et al. (2014), there is no combination of filters that will not bias the SFS for low and medium depth sequencing. The alternative is not to call SNPs and genotypes but instead to model the uncertainty of the data and to sum over all the possible genotypes without making a decision about any single genotype (Nielsen et al., 2012; Korneliussen et al., 2014) (Fig. 2B). This can be achieved by basing the analysis on so-called genotype likelihoods, which tries to capture all relevant information about the sequencing data for each site. This approach has been

Table 2
Tools for mapping short reads. For each tool, features are reported as in the publication and/or manual specifications.

Program	Algorithm type	Gapped alignment	PE	Read length (bp)	Indel size	Mapping quality	Color space	Note
ELANDev2 (Bauer et al., 2010)	Hash table of the reads (slower but more sensitive; usually has larger memory requirements).	Yes	Yes	15–32	1–10 bp	No	No	Designed for mapping Illumina reads (part of the CASAVA 1.8.2 pipeline). It is possible to map reads longer than 32 with additional scripts.
MAQ (Li et al., 2008a)		No	Yes	4–127	–	Yes	Yes	Gapped alignment is only available for paired-end reads, where only one read has indels (Li and Homer, 2010).
SHRiMP2 (David et al., 2011)		Yes	Yes	30–1000	Defined by the user	No	Yes	Originally developed to map in color space, i.e., SOLiD reads. Good option for mapping to a more distantly related species (allows a larger number of polymorphisms). Calculates a score for ranking alignments analogous to a mapping quality. Maintenance stopped as of 2014.
SOAP (Li et al., 2008b)	Hash table of the genome (slower but more sensitive; usually has larger memory requirements).	Yes	Yes	7–60	1–3 bp	No	No	In paired-end mode, gaps are allowed only in one read of each pair.
BFAST (Homer et al., 2009)		Yes	Yes	25–100	1–10 bp	No	Yes	High memory requirement.
Novoalign.v3 (14)		Yes	Yes	30–950	Up to 60% of the read length	Yes	Yes	High sensitivity (Li and Homer, 2010). Not freely available. Not open-source. High RAM requirements. NovoalignCS for color space mapping.
Mosaik (Lee et al., 2014)		Yes	Yes	15–1000	1–15 bp	Yes	Yes	Uses a neural-based training scheme to calculate mapping quality scores (default calibrated in the human genome). Able to map long reads from PacBio and Ion Torrent.
Stampy (Lunter and Goodson, 2011)		Yes	Yes	4–4000	1–15 bp	Yes	No	Recommended to be used in an hybrid mode with BWA. Uses a probabilistic model to calculate mapping qualities.
BSMAP (Xi and Li, 2009)		Yes	Yes	8–144	1–3 bp	No	No	Developed to map reads from bisulfite-treatment experiments.
Bowtie (Langmead et al., 2009)	Prefix/Suffix trie & BWT (faster but less sensitive).	No	No	4–1024	–	No	No	No length limit, but developed for short reads (~50 bp)
Bowtie2 (Langmead and Salzberg, 2012)		Yes	Yes	4–1000	Defined by the user	Yes	No	Incorporates a dynamic algorithm (swsse2) originally developed for protein alignment in order to find the optimal alignment between the reads and the reference sequence.
BWA aln (Li and Durbin, 2009)		Yes	Yes	4–200	Defined by the user	Yes	No	Able to map reads longer than 100 bp, but running times increase with length and error rate. Designed to find small indels.
BWA sw (Li and Durbin, 2010)		Yes	Yes	70-1 × 10^6	Defined by the user	Yes	No	Good for long reads with high indel/mismatch rates.
BWA mem (Li, 2013)		Yes	Yes	70-1 × 10^6	Defined by the user	Yes	No	Best performance with 70–100 bp reads.
SOAP2 (Li et al., 2009)		Yes	Yes	7–1000	Defined by the user	Yes	No	In paired-end mode, gaps are allowed only in one read of each pair.
GSNAP (Wu and Nacu, 2010)		Yes	Yes	14–no limit	1–9/1–30 bp	Yes	No	Developed to map reads from bisulfite-treatment experiments. Able to incorporate known exon–intron boundaries for mapping at splicing sites.

PE: supports mapping for paired-end data. Read length: supported read length. Indel size: gap(s) size range allowed in gapped alignments. Mapping quality: computes the probability of reporting the true alignment for a mapped read. Color space: supports mapping in color space. BWT — Burrows–Wheeler Transform; PE — Paired end.

successfully used in many other methods including estimation of allele frequencies (Kim et al., 2011), admixture proportion (Skotte et al., 2013), inbreeding coefficients (Vieira et al., 2013) and many other useful analysis (see (O'Rawe et al., 2015)) for a detailed review).

11. Online resources for training, problem solving and data sharing

At this point in this review, it has perhaps become obvious that genomics-based projects involve the efficient combination of many different types of software. However, it is sometimes difficult for a beginner to decide between different options of the same program, or choose the best approach. Online forums such as SEQanswers (http://seqanswers.com/) and Biostars (https://www.biostars.org/) provide useful information and guidance. Here, other software users share their experiences and one can easily find answer to questions of all levels of technical difficulty. Some websites follow a tutorial-like structure, such as the ANGUS website (angus.readthedocs.org/en/2015), which provides a compilation of lectures on various topics in NGS. Furthermore, a more comprehensive view on workflows in biological computing and the importance of reproducibility can be found in Shade and Teal (2015).

The sequence data should be made available to the scientific community, and uploaded into public databases such as the Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) or the European Nucleotide Archive (http://www.ebi.ac.uk/ena). Files that contribute to reproducibility in research such as scripts used during data analysis, intermediate data files, and phylogenetic trees can also be made available in various ways (Whitlock, 2011).

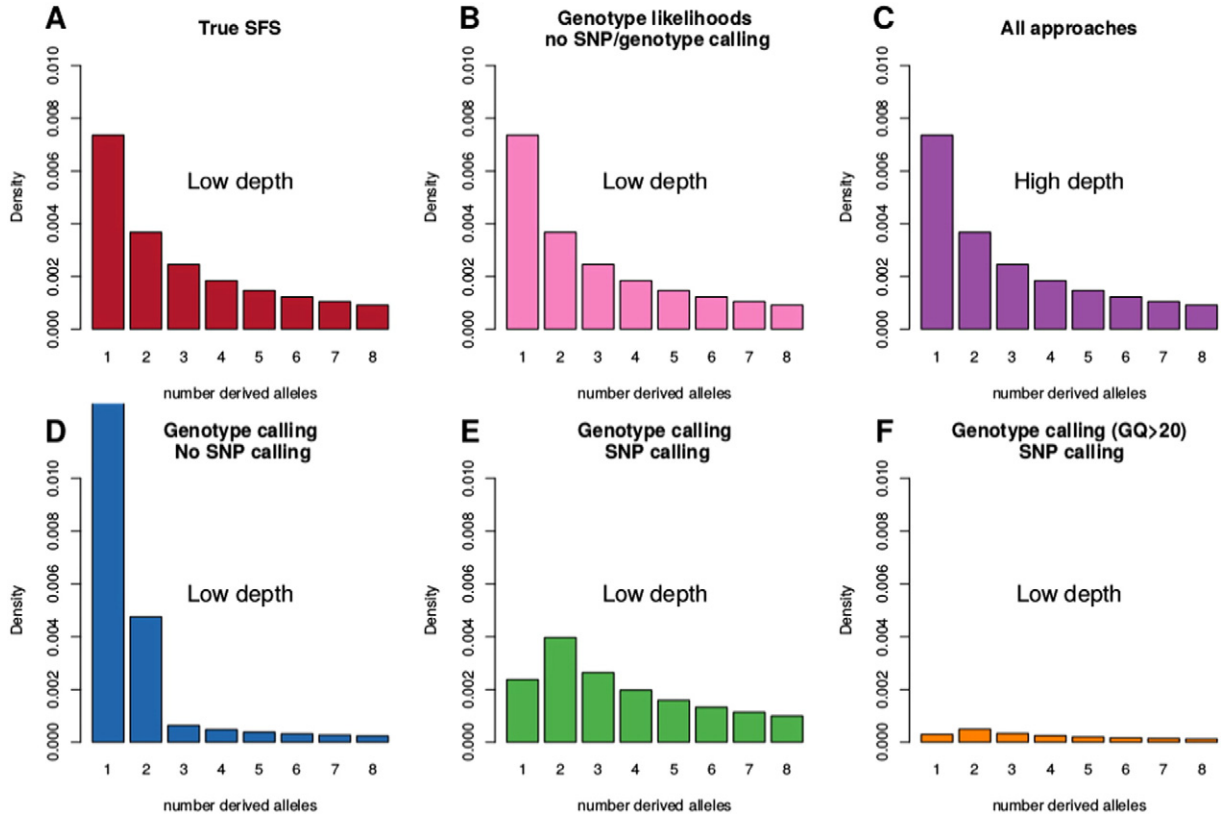

Fig. 2. Effect of SNP and genotype calling on the site frequency spectrum (SFS). The figure is a qualitative summary of the results from Nielsen et al. (2011, 2012); Han et al. (2014); Nevado et al. (2014)) which are based on both simulations and real data of low depth (<5×) and high depth (>10×) data. **A**: the True SFS; **B**: the estimated SFS based on genotype likelihoods; **C**: the SFS estimated with any of the approaches when the sequencing depth is high; **D**: the estimated SFS based on called genotypes without joint SNP calling; **E**: SFS from called genotypes after SNP calling; **F**: SFS from SNP and genotype calling with a cutoff based on the genotype quality.

12. Concluding remarks

Evolutionary biology is undergoing an exciting transition with the possibilities presented by high-throughput sequencing. However, it is key to choose the combinations of laboratory procedures and sequencing approaches that optimize the data for addressing a specific biological question. Most of the available sequencing and analysis approaches are designed for high quality genomic datasets from model organisms, and so extra care is necessary when applying them to limited data sets. A good knowledge of the theoretical basis behind the methods is required for making the appropriate choices of parameters, which can be very project-specific. Furthermore, the classical SNP and genotype calling approaches are often inadequate and overlook a lot of useful information in the data. To take full advantage of the limited datasets characteristic of projects involving non-model organisms, a more than basic bioinformatics background is required of the biological researcher, as a considerable amount of effort must be put in obtaining an unbiased and representative dataset ready for an optimal biologically relevant analysis. Rushing into adopting new and improved NGS methods while disregarding the limitations discussed here may in fact be the largest current obstacle in answering key questions in evolutionary biology. In almost any case, sequencing methods will supply large amounts of data, and analytical methods will provide summary statistics that can be interpreted with more or less creativity. But only by being aware of the options and limitations at every stage of the NGS-based study can we truly leverage on the advantages of having a large dataset for a non-model organism and address questions that cannot be answered with model species.

Acknowledgments

RF is supported by Young Investigator grant VKR023446 from Villum Fonden. RH is supported by Young Investigator grant VKR023447 from Villum Fonden. RP has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 658706. MLZM is supported by Lundbeck Foundation grant R52-A5062.

References

Abu Almakarem, A.S., Heilman, K.L., Conger, H.L., Shtarkman, Y.M., Rogers, S.O., 2012. Extraction of DNA from plant and fungus tissues in situ. BMC Res. Notes 5. BioMed Central, p. 266 ([Internet], [cited 2016 Mar 18], Available from: http://bmcresnotes.biomedcentral.com/articles/10.1186/1756-0500-5-266).

Alkan, C., Sajjadian, S., Eichler, E.E., 2010. Limitations of next-generation genome sequence assembly. Nat. Methods 8. Nature Publishing Group, a division of Macmillan Publishers Limited, pp. 61–65 All Rights Reserved.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., et al., 2000. Gene ontology: tool for the unification of biology. Nat. Genet. 25, 25–29. http://dx.doi.org/10.1038/75556 ([Internet], Available from).

Attwood, T.K., Bradley, P., Flower, D.R., Gaulton, A., Maudling, N., Mitchell, A.L., et al., 2003. PRINTS and its automatic supplement, prePRINTS. Nucleic Acids Res. 31, 400–402 ([Internet], [cited 2016 Feb 15], Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=165477&tool=pmcentrez&rendertype=abstract).

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., et al., 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS one 3. Public Library of Science, p. e3376 ([Internet], [cited 2014 Jul 9], Available from: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0003376).

Baker, M., 2012. De novo genome assembly: what every biologist should know. Nat. Methods 9. Nature Publishing Group, a division of Macmillan Publishers Limited, pp. 333–337. http://dx.doi.org/10.1038/nmeth.1935 ([Internet]. All Rights Reserved, [cited 2014 Dec 17], Available from:).

Barshis, D.J., Ladner, J.T., Oliver, T.A., Seneca, F.O., Traylor-Knowles, N., Palumbi, S.R., 2013. Genomic basis for coral resilience to climate change. Proc. Natl. Acad. Sci. U. S. A. 110,

1387–1392 ([Internet], [cited 2015 Aug 27], Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3557039&tool=pmcentrez&rendertype=abstract).
Bauer, M.J., Cox, A.J., Evers, D.J., 2010. ELANDv2 - fast gapped read mapping for illumina reads. Proceeding of the 18th Annual Conference on Intelligent Systems for Molecular Biology, J04. ISCB.
Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., et al., 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456, 53–59.
Berlin, K., Koren, S., Chin, C.-S., Drake, J.P., Landolin, J.M., Phillippy, A.M., 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat. Biotechnol. 33, 623–630.
Bertolini, F., Scimone, C., Geraci, C., Schiavo, G., Utzeri, V.J., Chiofalo, V., et al., 2015. Next generation semiconductor based sequencing of the donkey (Equus asinus) genome provided comparative sequence data against the horse genome and a few millions of single nucleotide polymorphisms. In: te Pas, M.F.W. (Ed.)PLoS one 10, p. e0131925.
Bi, K., Linderoth, T., Vanderpool, D., Good, J.M., Nielsen, R., Moritz, C., 2013. Unlocking the vault: next-generation museum population genomics. Mol. Ecol. 22, 6018–6032 ([Internet], [cited 2015 Aug 27], Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4134471&tool=pmcentrez&rendertype=abstract).
Bradnam, K.R., Fass, J.N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., et al., 2013. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. Gigascience 2, 10.
Brejová, B., Vinar, T., Chen, Y., Wang, S., Zhao, G., Brown, D.G., et al., 2009. Finding genes in *Schistosoma japonicum*: annotating novel genomes with help of extrinsic evidence. Nucleic Acids Res. 37, e52 ([Internet], [cited 2016 Mar 10], Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2673418&tool=pmcentrez&rendertype=abstract).
Brendel, V., Xing, L., Zhu, W., 2004. Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus. Bioinformatics 20, 1157–1169 ([Internet], [cited 2016 Mar 11], Available from: http://www.ncbi.nlm.nih.gov/pubmed/14764557).
Buermans, H.P.J., den Dunnen, J.T., 1842. Next generation sequencing technology: advances and applications. Biochim. Biophys. Acta Mol. Basis Dis. 2014, 1932–1941.
Burrows, M., Wheeler, D., 1994. A block-sorting lossless data compression algorithm. Tech. Rep. 124. CA Digit. Equip. Corp., Palo Alto
Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al., 2009. BLAST+: architecture and applications. BMC Bioinf. 10, 421 ([Internet], [cited 2014 Jul 9], Available from:) http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2803857&tool=pmcentrez&rendertype=abstract).
Campos, P.F., Willerslev, E., Gilbert, M.T.P., 2009. Isolation of DNA from ancient samples. In: Liu, D. (Ed.), Handb. Nucleic Acid Purif.Taylor & Francis, pp. 441–461
Carneiro, M., Albert, F.W., Afonso, S., Pereira, R.J., Burbano, H., Campos, R., et al., 2014. The genomic architecture of population divergence between subspecies of the European rabbit. PLoS Genet 10. Public Library of Science, p. e1003519 ([Internet], [cited 2015 Oct 5], Available from: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003519).
Chang, Z., Li, G., Liu, J., Zhang, Y., Ashby, C., Liu, D., et al., 2015. Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. Genome Biol. 16. BioMed Central Ltd., p. 30 ([Internet], [cited 2015 Nov 27], Available from: http://genomebiology.com/2015/16/1/30)
Cherf, G.M., Lieberman, K.R., Rashid, H., Lam, C.E., Karplus, K., Akeson, M., 2012. Automated forward and reverse ratcheting of DNA in a nanopore at 5-Å precision. Nat. Biotechnol. 30, 344–348.
Chia, J.J.-M., Song, C., Bradbury, P.J.P., Costich, D., de Leon, N., Doebley, J., et al., 2012. Maize HapMap2 identifies extant variation from a genome in flux. Nat. Genet. 44 (803–U238. [Internet] [cited 2013 May 3], Available from: http://www.nature.com/ng/journal/v44/n7/abs/ng.2313.html).
Collins, F.S., Weissman, S.M., 1984. Directional cloning of DNA fragments at a large distance from an initial probe: a circularization method. Proc. Natl. Acad. Sci. 81, 6812–6816.
Compeau, P.E.C., Pevzner, P.A., Tesler, G., 2011. How to apply de Bruijn graphs to genome assembly. Nat. Biotechnol. 29. Nature Publishing Group, a division of Macmillan Publishers Limited, pp. 987–991 All Rights Reserved
Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., Robles, M., 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21, 3674–3676 ([Internet], [cited 2014 Jul 9], Available from: http://bioinformatics.oxfordjournals.org/content/21/18/3674.full).
da Fonseca, R.R., Smith, B.D., Wales, N., Cappellini, E., Skoglund, P., Fumagalli, M., et al., 2015. The origin and evolution of maize in the American Southwest. Nat. Plants. Nature Publishing Group, pp. 1–5 ([Internet], [cited 2015 Jan 8], in press, Available from: http://www.nature.com/articles/nplants20143).
Davey, J.W., Cezard, T., Fuentes-Utrilla, P., Eland, C., Gharbi, K., Blaxter, M.L., 2013. Special features of RAD sequencing data: implications for genotyping. Mol. Ecol. 22, 3151–3164 ([Internet], [cited 2015 Nov 26], Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3712469&tool=pmcentrez&rendertype=abstract).
David, M., Dzamba, M., Lister, D., Ilie, L., Brudno, M., 2011. SHRiMP2: sensitive yet practical SHort read mapping. Bioinformatics 27, 1011–1012 ([Internet], [cited 2016 Jan 29], Available from: http://www.ncbi.nlm.nih.gov/pubmed/21278192).
Denoeud, F., Carretero-Paulet, L., Dereeper, A., Droc, G., Guyot, R., Pietrella, M., et al., 2014. The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science 345, 1181–1184 ((80-.). [Internet] [cited 2014 Sep 5], Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.1255274).
Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al., 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 ([Internet], [cited 2014 Jul 13], Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3530905&tool=pmcentrez&rendertype=abstract).
du Buy, H.G., Riley, F.L., 1967. Hybridization between the nuclear and kinetoplast DNA's of *Leishmania enrietti* and between nuclear and mitochondrial DNA's of mouse liver. Proc. Natl. Acad. Sci. U. S. A. 57, 790–797.
Dunn, C.W., Giribet, G., Edgecombe, G.D., Hejnol, A., 2014. Animal phylogeny and its evolutionary implications*. Annu. Rev. Ecol. Evol. Syst. 45, 371–395 ([Internet]. Annual Reviews, [cited 2015 Nov 19], Available from: http://www.annualreviews.org/doi/abs/10.1146/annurev-ecolsys-120213-091627).
Earl, D., Bradnam, K., St John, J., Darling, A., Lin, D., Fass, J., et al., 2011. Assemblathon 1: a competitive assessment of de novo short read assembly methods. Genome Res. 21, 2224–2241 ([Internet], [cited 2013 Feb 28], Available from:) http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3227110&tool=pmcentrez&rendertype=abstract).
Eaton, D.A.R., Hipp, A.L., González-Rodríguez, A., Cavender-Bares, J., Gon Alez-Rodríguez, A., Cavender-Bares, J., 2015. Historical introgression among the American live oaks and the comparative nature of tests for introgression. Evolution 69, 2587–2601 ([Internet], [cited 2015 Aug 26], Available from: http://www.ncbi.nlm.nih.gov/pubmed/26299374).
Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., et al., 2009. Real-time DNA sequencing from single polymerase molecules. Science 323, 133–138 (80-.).
Espregueira Themudo, G., Babik, W., Arntzen, J.W., 2009. A combination of techniques proves useful in the development of nuclear markers in the newt genus *Triturus*. Mol. Ecol. Resour. 9. Wiley-Blackwell Publishing, pp. 1160–1162
Faircloth, B.C., McCormack, J.E., Crawford, N.G., Harvey, M.G., Brumfield, R.T., Glenn, T.C., 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst. Biol. 61, 717–726 ([Internet], [cited 2015 Oct 1], Available from: http://sysbio.oxfordjournals.org/content/61/5/717.abstract).
Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, S.J., Hotz, H.-R., et al., 2008. The Pfam protein families database. Nucleic Acids Res. 36, D281–D288 ([Internet], [cited 2013 Mar 4], Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2238907&tool=pmcentrez&rendertype=abstract).
Florea, L., Souvorov, A., Kalbfleisch, T.S., Salzberg, S.L., 2011. Genome assembly has a major impact on gene content: a comparison of annotation in two *Bos taurus* assemblies. PLoS ONE 6, e21400.
Gautier, M., Gharbi, K., Cezard, T., Foucaud, J., Kerdelhué, C., Pudlo, P., et al., 2013. The effect of RAD allele dropout on the estimation of genetic variation within and between populations. Mol. Ecol. 22, 3165–3178 ([Internet], [cited 2015 Nov 26], Available from: http://www.ncbi.nlm.nih.gov/pubmed/23110526).
Gayral, P., Melo-Ferreira, J., Glémin, S., Bierne, N., Carneiro, M., Nabholz, B., et al., 2013. Reference-free population genomics from next-generation transcriptome data and the vertebrate–invertebrate gap. In: Welch, J.J. (Ed.), PLoS Genet 9. Public Library of Science, p. e1003457 ([Internet], [cited 2013 Apr 12], Available from: http://dx.plos.org/10.1371/journal.pgen.1003457).
Geng, C., Chen, Y., Wu, K., Cai, Q., Wang, Y., Lang, Y., et al., 2012. Paired-end sequencing of long-range DNA fragments for de novo assembly of large, complex mammalian genomes by direct intra-molecule ligation. In: Aboobaker, A.A. (Ed.), PLoS one. 7, p. e46211.
Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., et al., 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. U. S. A. 108, 1513–1518.
Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., et al., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29. Nature Publishing Group, a division of Macmillan Publishers Limited, pp. 644–652 ([Internet], All Rights Reserved, [cited 2013 Feb 27], Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3571712&tool=pmcentrez&rendertype=abstract)
Grover, C.E., Salmon, A., Wendel, J.F., 2012. Targeted sequence capture as a powerful tool for evolutionary analysis. Am. J. Bot. 99, 312–319 ([Internet], [cited 2015 Nov 26], Available from: http://www.amjbot.org/content/99/2/312.long).
Gurevich, A., Saveliev, V., Vyahhi, N., Tesler, G., 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29, 1072–1075 ([Internet], [cited 2014 Jul 10], Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3624806&tool=pmcentrez&rendertype=abstract).
Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H., Bustamante, C.D., 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. In: McVean, G. (Ed.)PLoS Genet 5. Public Library of Science, e1000695. http://dx.doi.org/10.1371/journal.pgen.1000695 ([Internet], [cited 2013 May 23], Available from:).
Haas, B.J., Zeng, Q., Pearson, M.D., Cuomo, C.A., Wortman, J.R., 2011. Approaches to fungal genome annotation. Mycology 2, 118–141 ([Internet], [cited 2016 Mar 2], Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3207268&tool=pmcentrez&rendertype=abstract).
Han, E., Sinsheimer, J.S., Novembre, J., 2014. Characterizing bias in population genetic inferences from low-coverage sequencing data. Mol. Biol. Evol. 31, 723–735 ([Internet], [cited 2015 Nov 26], Available from: http://mbe.oxfordjournals.org/content/early/2013/11/27/molbev.mst229.abstract).
Hargreaves, A.D., Mulley, J.F., 2015. Assessing the utility of the Oxford Nanopore MinION for snake venom gland cDNA sequencing. PeerJ 3. PeerJ Inc., p. e1441 ([Internet], [cited 2015 Nov 30], Available from: https://peerj.com/articles/1441).
Hayer KE, Pizarro A, Lahens NF, Hogenesch JB, Grant, GR. Benchmark analysis of algorithms for determining and quantifying full-length mRNA splice forms from RNA-seq data. Bioinformatics. 2015 ([Internet]. [cited 2015 Sep 10];btv488 –, Available from: http://bioinformatics.oxfordjournals.org/content/early/2015/09/03/bioinformatics.btv488).

Heyduk, K., Trapnell, D.W., Barrett, C.F., Leebens-Mack, J., 2016. Phylogenomic analyses of species relationships in the genus *Sabal* (Arecaceae) using targeted sequence capture. Biol. J. Linn. Soc. 117, 106–120 ([Internet], [cited 2016 Mar 18], Available from: http://doi.wiley.com/10.1111/bij.12551).

Hohenlohe, P.A., Bassham, S., Etter, P.D., Stiffler, N., Johnson, E.A., Cresko, W.A., 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet. 6. Public Library of Science, p. e1000862 ([Internet], [cited 2015 May 13], Available from: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000862)

Homer, N., Merriman, B., Nelson, S.F., 2009. BFAST: an alignment tool for large scale genome resequencing. In: Creighton, C. (Ed.)PLoS one 4, p. e7767.

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M.C., et al., 2015. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 44, D286–D293 ([Internet], [cited 2015 Nov 20], Available from: http://nar.oxfordjournals.org/content/44/D1/D286).

Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M., Otto, T.D., 2013. REAPR: a universal tool for genome assembly evaluation. Genome Biol. 14, R47 ([Internet]. BioMed Central, [cited 2015 Dec 10], Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2013-14-5-r47).

Keller, O., Odronitz, F., Stanke, M., Kollmar, M., Waack, S., 2008. Scipio: using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species. BMC Bioinf. 9. BioMed Central, p. 278 ([Internet], [cited 2016 Mar 10], Available from:) http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-278)

Kent, W.J., 2002. BLAT—the BLAST-like alignment tool. Genome Res. 12, 656–664 ([Internet], [cited 2014 Dec 20], Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=187518&tool=pmcentrez&rendertype=abstract).

Kim, S.Y., Lohmueller, K.E., Albrechtsen, A., Li, Y., Korneliussen, T., Tian, G., et al., 2011. Estimation of allele frequency and association mapping using next-generation sequencing data. BMC Bioinf. 12. BioMed Central Ltd., p. 231 ([Internet], [cited 2015 Nov 26], Available from: http://www.biomedcentral.com/1471-2105/12/231)

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., Salzberg, S.L., 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14, R36 ([Internet], [cited 2014 Jul 9], Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4053844&tool=pmcentrez&rendertype=abstract).

Kim, D., Langmead, B., Salzberg, S.L., 2015. HISAT: a fast spliced aligner with low memory requirements. Nat. Methods 12. Nature Publishing Group, a Division of Macmillan Publishers Limited, pp. 357–360. http://dx.doi.org/10.1038/nmeth.3317 ([Internet], All Rights Reserved, [cited 2015 Mar 11]. Available from).

Korf, I., 2004. Gene finding in novel genomes. BMC Bioinf. 5, 59 (BioMed Central, [Internet], [cited 2016 Feb 17], Available from: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-5-59).

Korneliussen, T.S., Albrechtsen, A., Nielsen, R., 2014. ANGSD: analysis of next generation sequencing data. BMC Bioinf. 15. BioMed Central Ltd., p. 356 ([Internet], [cited 2015 Jul 31], Available from: http://www.biomedcentral.com/1471-2105/15/356)

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359.

Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25 ([Internet], [cited 2014 Jul 9], Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2690996&tool=pmcentrez&rendertype=abstract).

Lee, W.-P., Stromberg, M.P., Ward, A., Stewart, C., Garrison, E.P., Marth, G.T., 2014. MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. In: Hsiao, C.K. (Ed.), PLoS one 9. Public Library of Science, p. e90581.

Lemmon, A.R., Emme, S.A., Lemmon, E.M., 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst. Biol. 61, 727–744 ([Internet], [cited 2015 Oct 13], Available from: http://sysbio.oxfordjournals.org/content/61/5/727.abstract).

Letunic, I., Doerks, T., Bork, P., 2015. SMART: recent updates, new developments and status in 2015. Nucleic Acids Res. 43, D257–D260 ([Internet], [cited 2015 Mar 22], Available from: http://nar.oxfordjournals.org/content/43/D1/D257.abstract?keytype=ref&ijkey=gOFbK1RrNpOsExA).

Li, H., 2013. Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM. p. 3 ([cited 2015 Nov 30], Available from: http://arxiv.org/abs/1303.3997).

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754–1760 ([Internet], [cited 2014 Jan 20], Available from: http://bioinformatics.oxfordjournals.org/content/25/14/1754.long).

Li, H., Durbin, R., 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics 26, 589–595.

Li, H., Homer, N., 2010. A survey of sequence alignment algorithms for next-generation sequencing. Brief. Bioinform. 11, 473–483.

Li, H., Ruan, J., Durbin, R., 2008a. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 18, 1851–1858 ([Internet], [cited 2014 Jul 9], Available from: http://genome.cshlp.org/content/early/2008/01/01/gr.078212.108).

Li, R., Li, Y., Kristiansen, K., Wang, J., 2008b. SOAP: short oligonucleotide alignment program. Bioinformatics 24, 713–714.

Li, R., Yu, C., Li, Y., Lam, T.-W., Yiu, S.-M., Kristiansen, K., et al., 2009. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25, 1966–1967.

Li, B., Fillmore, N., Bai, Y., Collins, M., Thomson, J.A., Stewart, R., et al., 2014. Evaluation of de novo transcriptome assemblies from RNA-Seq data. Genome Biol. 15. BioMed Central Ltd, p. 553 ([Internet] [cited 2015 Jul 20] Available from: http://genomebiology.com/2014/15/12/553)

Liu, X., Fu, Y.-X., 2015. Corrigendum: Exploring population size changes using SNP frequency spectra. Nat. Genet. 47. Nature Publishing Group, a Division of Macmillan Publishers Limited, p. 1099. http://dx.doi.org/10.1038/ng0915-1099a ([Internet], All Rights Reserved, [cited 2015 Nov 26], Available from:)

Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15. BioMed Central, p. 550 ([Internet], [cited 2016 Jan 31], Available from: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8)

Lunter, G., Goodson, M., 2011. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. 21, 936–939.

M., R., 1998. DNA SEQUENCING: a sequencing method based on real-time pyrophosphate. Science 281, 363–365 (80-.).

Madoui, M.-A., Engelen, S., Cruaud, C., Belser, C., Bertrand, L., Alberti, A., et al., 2015. Genome assembly using nanopore-guided long and error-free DNA reads. BMC Genomics 16.

Mamanova, L., Coffey, A.J., Scott, C.E., Kozarewa, I., Turner, E.H., Kumar, A., et al., 2010. Target-enrichment strategies for next-generation sequencing. Nat. Methods 7. Nature Publishing Group, pp. 111–118. http://dx.doi.org/10.1038/nmeth.1419 ([Internet], [cited 2015 Mar 9], Available from).

Maretty, L., Sibbesen, J.A., Krogh, A., 2014. Bayesian transcriptome assembly. Genome Biol. 15. BioMed Central, p. 501 ([Internet], [cited 2016 Mar 10], Available from: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0501-4)

McCormack, J.E., Faircloth, B.C., Crawford, N.G., Gowaty, P.A., Brumfield, R.T., Glenn, T.C., 2012. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. Genome Res. 22, 746–754 ([Internet], [cited 2015 Nov 29], Available from: http://genome.cshlp.org/content/22/4/746.abstract).

McKernan, K.J., Peckham, H.E., Costa, G.L., McLaughlin, S.F., Fu, Y., Tsung, E.F., et al., 2009. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res. 19, 1527–1541.

Merriman, B., R&D Team IT, Rothberg, J.M., 2012. Progress in Ion Torrent semiconductor chip based sequencing: nanoanalysis. Electrophoresis 33, 3397–3417.

Mitchell, A., Chang, H.-Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., et al., 2014. The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res. 43, D213–D221 ([Internet], [cited 2014 Nov 28], Available from: http://nar.oxfordjournals.org/content/43/D1/D213).

Musacchia, F., Basu, S., Petrosino, G., Salvemini, M., Sanges, R., 2015. Annocript: a flexible pipeline for the annotation of transcriptomes able to identify putative long noncoding RNAs. Bioinformatics 31, 2199–2201 ([Internet] [cited 2016 Mar 14], Available from: http://bioinformatics.oxfordjournals.org/content/early/2015/02/19/bioinformatics.btv106).

Nadeau, N.J., Whibley, A., Jones, R.T., Davey, J.W., Dasmahapatra, K.K., Baxter, S.W., et al., 2012. Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci. 367, 343–353 ([Internet], [cited 2015 Nov 26], Available from: http://rstb.royalsocietypublishing.org/content/367/1587/343.long).

Nevado, B., Ramos-Onsins, S.E., Perez-Enciso, M., 2014. Resequencing studies of nonmodel organisms using closely related reference genomes: optimal experimental designs and bioinformatics approaches for population genomics. Mol. Ecol. 23, 1764–1779 ([Internet], [cited 2015 Oct 28], Available from: http://doi.wiley.com/10.1111/mec.12693).

Nicholls, J.A., Pennington, R.T., Koenen, E.J.M., Hughes, C.E., Hearn, J., Bunnefeld, L., et al., 2015. Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the neotropical rain forest genus Inga (Leguminosae: Mimosoideae). Front. Plant Sci. 6. Frontiers, p. 710 ([Internet], [cited 2015 Nov 27], Available from: http://journal.frontiersin.org/article/10.3389/fpls.2015.00710/abstract)

Nielsen, R., Paul, J.S., Albrechtsen, A., Song, Y.S., 2011. Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. 12. Nature Publishing Group, a Division of Macmillan Publishers Limited, pp. 443–451. http://dx.doi.org/10.1038/nrg2986 ([Internet], All Rights Reserved, [cited 2014 Jul 9], Available from:)

Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., Wang, J., 2012. SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. PLoS one 7. Public Library of Science, p. e37558 ([Internet], Available from: 10.1371%2Fjournal.pone.0037558).

O'Neil, S.T., Emrich, S.J., 2013. Assessing de novo transcriptome assembly metrics for consistency and utility. BMC Genomics 14, 465 ([Internet] [cited 2016 Feb 17], Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3733778&tool=pmcentrez&rendertype=abstract).

O'Rawe, J.A., Ferson, S., Lyon, G.J., 2015. Accounting for uncertainty in DNA sequencing data. Trends Genet. 31. Elsevier, pp. 61–66 ([Internet], [cited 2015 Nov 26], Available from: http://www.cell.com/article/S0168952514002091/fulltext)

Parra, G., Bradnam, K., Korf, I., 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23, 1061–1067 ([Internet] [cited 2014 Jul 11], Available from: http://bioinformatics.oxfordjournals.org/content/23/9/1061.abstract).

Pereira, R.J., Barreto, F.S., Pierce, T., Carneiro, M., Burton, R.S., 2016. Transcriptome-wide patterns of divergence during allopatric evolution. Mol. Ecol.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T., Salzberg, S.L., 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. 33. Nature Publishing Group, a Division of Macmillan Publishers Limited, pp. 290–295. http://dx.doi.org/10.1038/nbt.3122 ([Internet], All Rights Reserved, [cited 2015 Feb 18]. Available from)

Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S., Hoekstra, H.E., 2012. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS one 7. Public Library of Science, p. e37135

([Internet], [cited 2014 Jul 11], Available from:) http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0037135).
Piskol, R., Ramaswami, G., Li, J.B., 2013. Reliable identification of genomic variants from RNA-seq data. Am. J. Hum. Genet. 93, 641–651 ([Internet], [cited 2016 Jan 20], Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3791257&tool=pmcentrez&rendertype=abstract).
Prum, R.O., Berv, J.S., Dornburg, A., Field, D.J., Townsend, J.P., Lemmon, E.M., et al., 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature 526. Nature Publishing Group, a Division of Macmillan Publishers Limited, pp. 569–573 ([Internet], All Rights Reserved; [cited 2015 Oct 7], Available from: http://www.nature.com/doifinder/10.1038/nature15697).
Putnam, N.H., O'Connell, B., Stites, J.C., Rice, B.J., Fields, A., Hartley, P.D., et al., 2015. Chromosome-Scale Shotgun Assembly Using an In Vitro Method for Long-Range Linkage.
Quail, M., Smith, M.E., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., et al., 2012. A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. BMC Genomics 13, 341.
Renaut, S., Grassa, C.J., Yeaman, S., Moyers, B.T., Lai, Z., Kane, N.C., et al., 2013. Genomic islands of divergence are not affected by geography of speciation in sunflowers. Nat. Commun. 4. Nature Publishing Group, a Division of Macmillan Publishers Limited, p. 1827. http://dx.doi.org/10.1038/ncomms2833 ([Internet], All Rights Reserved, [cited 2015 Aug 30], Available from)
Romiguier, J., Gayral, P., Ballenghien, M., Bernard, A., Cahais, V., Chenuil, A., et al., 2014. Comparative population genomics in animals uncovers the determinants of genetic diversity. Nature 515. Nature Publishing Group, a Division of Macmillan Publishers Limited, pp. 261–263 ([Internet], All Rights Reserved, [cited 2014 Aug 20], Available from: http://www.nature.com/doifinder/10.1038/nature13685).
Salzberg, S.L., Yorke, J.A., 2005. Beware of mis-assembled genomes. Bioinformatics 21, 4320–4321 ([Internet] [cited 2016 Mar 10]; Available from: http://bioinformatics.oxfordjournals.org/content/21/24/4320.full).
Salzberg, S.L., Phillippy, A.M., Zimin, A., Puiu, D., Magoc, T., Koren, S., et al., 2012. GAGE: a critical evaluation of genome assemblies and assembly algorithms. Genome Res. 22, 557–567 ([Internet], [cited 2015 Dec 1], Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3290791&tool=pmcentrez&rendertype=abstract).
Samaniego Castruita, J.A., Zepeda Mendoza, M.L., Barnett, R., Wales, N., Gilbert, M.T.P., 2015. Odintifier — a computational method for identifying insertions of organellar origin from modern and ancient high-throughput sequencing data based on haplotype phasing. BMC Bioinf. 16, 232.
Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., et al., 1977. Nucleotide sequence of bacteriophage φX174 DNA. Nature 265, 687–695.
Schbath, S., Martin, V., Zytnicki, M., Fayolle, J., Loux, V., Gibrat, J.-F., 2012. Mapping reads on a genomic sequence: an algorithmic overview and a practical comparative analysis. J. Comput. Biol. 19, 796–813.
Schirmer, M., Ijaz, U.Z., D'Amore, R., Hall, N., Sloan, W.T., Quince, C., 2015. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res. 43, e37.
Schulz, M.H., Zerbino, D.R., Vingron, M., 2012. Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28, 1086–1092 ([Internet] [cited 2015 Jun 22], Available from: http://bioinformatics.oxfordjournals.org/content/28/8/1086.abstract).
Servant, F., Bru, C., Carrère, S., Courcelle, E., Gouzy, J., Peyruc, D., et al., 2002. ProDom: automated clustering of homologous domains. Brief. Bioinform. 3, 246–251 ([Internet], [cited 2016 Mar 10], Available from: http://www.ncbi.nlm.nih.gov/pubmed/12230033).
Shade, A., Teal, T.K., 2015. computing workflows for biologists: a roadmap. PLoS Biol. 13. Public Library of Science, p. e1002303 ([Internet], [cited 2015 Nov 24], Available from: http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002303)
Shao, W., Boltz, V.F., Spindler, J.E., Kearney, M.F., Maldarelli, F., Mellors, J.W., et al., 2013. Analysis of 454 sequencing error rate, error sources, and artifact recombination for detection of low-frequency drug resistance mutations in HIV-1 DNA. Retrovirology 10, 18.
Sigrist, C.J.A., Cerutti, L., de Castro, E., Langendijk-Genevaux, P.S., Bulliard, V., Bairoch, A., et al., 2010. PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res. 38, D161–D166 ([Internet], [cited 2013 Feb 28], Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2808866&tool=pmcentrez&rendertype=abstract).
Skotte L, Korneliussen TS, Albrechtsen A. Estimating individual admixture proportions from next generation sequencing data. Genetics [Internet]. 2013 ([cited 2013 Oct 26]; genetics.113.154138 – Available from: http://www.genetics.org/content/early/2013/09/03/genetics.113.154138.abstract?sid = a5f549bf-d0b5-407b-9c13-24142e579370)
Slater, G.S.C., Birney, E., 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinf. 6. BioMed Central, p. 31 ([Internet], [cited 2016 Feb 1], Available from: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-6-31)
Smith, S.A., Wilson, N.G., Goetz, F.E., Feehery, C., Andrade, S.C.S., Rouse, G.W., et al., 2011. Resolving the evolutionary relationships of molluscs with phylogenomic tools. Nature 480. Nature Publishing Group, a Division of Macmillan Publishers Limited, pp. 364–367. http://dx.doi.org/10.1038/nature10526 ([Internet], [cited 2015 Nov 24], All Rights Reserved. Available from).
Smith-Unna, RD, Boursnell, C, Patro, R, Hibberd, JM, Kelly, S. TransRate: reference free quality assessment of de-novo transcriptome assemblies. Cold Spring Harb. Labs J.; 2015. ([Internet]. bioRxiv. Jun, Available from: http://biorxiv.org/content/early/2015/06/27/021626.abstract).
Sodergren, E., Weinstock, G.M., Davidson, E.H., Cameron, R.A., Gibbs, R.A., Angerer, R.C., et al., 2006. The genome of the sea urchin *Strongylocentrotus purpuratus*. Science 314, 941–952 ([Internet] [cited 2016 Feb 9], Available from:) http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3159423&tool=pmcentrez&rendertype=abstract).
Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., Morgenstern, B., 2006. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 34, W435–W439 ([Internet], [cited 2016 Jan 1], Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1538822&tool=pmcentrez&rendertype=abstract).
Strong, M.J., Xu, G., Morici, L., Splinter Bon-Durant, S., Baddoo, M., Lin, Z., et al., 2014. Microbial contamination in next generation sequencing: implications for sequence-based analysis of clinical samples. PLoS Pathog. 10. Public Library of Science, p. e1004437 ([Internet]. [cited 2016 Mar 10], Available from: http://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1004437)
Todd, E.V., Black, M.A., Gemmell, N.J., 2016. The power and promise of RNA-seq in ecology and evolution. Mol. Ecol. 25 ([Internet], [cited 2016 Jan 12]:n/a – n/a. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26756714).
Toonen, R.J., Puritz, J.B., Forsman, Z.H., Whitney, J.L., Fernandez-Silva, I., Andrews, K.R., et al., 2013. ezRAD: a simplified method for genomic genotyping in non-model organisms. PeerJ 1, e203 ([Internet], [cited 2015 Nov 6], Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3840413&tool=pmcentrez&rendertype=abstract).
Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., et al., 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28. Nature Publishing Group, pp. 511–515 ([Internet], [cited 2014 Jul 9], Available from: http://www.nature.com/nbt/journal/v28/n5/pdf/nbt.1621.pdf)
Tsagkogeorga, G., Cahais, V., Galtier, N., 2012. The population genomics of a fast evolver: high levels of diversity, functional constraint, and molecular adaptation in the tunicate *Ciona intestinalis*. Genome Biol. Evol. 4, 740–749 ([Internet], [cited 2015 Nov 30], Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3509891&tool=pmcentrez&rendertype=abstract).
Ummat, A., Bashir, A., 2014. Resolving complex tandem repeats with long reads. Bioinformatics 30, 3491–3498.
van Dijk, E.L., Auger, H., Jaszczyszyn, Y., Thermes, C., 2014. Ten years of next-generation sequencing technology. Trends Genet. 30, 418–426.
Vieira FG, Fumagalli M, Albrechtsen A, Nielsen R. Estimating inbreeding coefficients from NGS data: impact on genotype calling and allele frequency estimation. Genome Res. 2013 ([Internet], [cited 2013 Oct 17]; gr.157388.113 –. Available from: http://genome.cshlp.org/content/early/2013/10/02/gr.157388.113.abstract)
Wagner, C.E., Keller, I., Wittwer, S., Selz, O.M., Mwaiko, S., Greuter, L., et al., 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Mol. Ecol. 22, 787–798 ([Internet], [cited 2015 Nov 21], Available from: http://www.ncbi.nlm.nih.gov/pubmed/23057853).
Wang, S., Meyer, E., McKay, J.K., Matz, M.V., 2012. 2b-RAD: a simple and flexible method for genome-wide genotyping. Nat. Methods 9. Nature Publishing Group, a Division of Macmillan Publishers Limited, pp. 808–810. http://dx.doi.org/10.1038/nmeth.2023 ([Internet], All Rights Reserved, [cited 2015 Nov 9], Available from).
Wang, M., Beck, C.R., English, A.C., Meng, Q., Buhay, C., Han, Y., et al., 2015. PacBio-LITS: a large-insert targeted sequencing method for characterization of human disease-associated chromosomal structural variations. BMC Genomics 16.
Whitlock, M.C., 2011. Data archiving in ecology and evolution: best practices. Trends Ecol. Evol. 26, 61–65 ([Internet], [cited 2016 Mar 18], Available from: http://www.sciencedirect.com/science/article/pii/S0169534710002697).
Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, et al. Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc. Natl. Acad. Sci. 2014 ([Internet], [cited 2014 Oct 30]; Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.1323926111)
Wong, P.B., Wiley, E.O., Johnson, W.E., Ryder, O.A., O'Brien, S.J., Haussler, D., et al., 2012. Tissue sampling methods and standards for vertebrate genomics. Gigascience. BioMed Central Ltd., pp. 1–8 ([Internet], [cited 2015 Nov 30], Available from: http://www.gigasciencejournal.com/content/1/1/8).
Wu, T.D., Nacu, S., 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26, 873–881.
Xi, Y., Li, W., 2009. BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinf. 10, 232.
Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., et al., 2014. SOAPdenovo-trans: de novo transcriptome assembly with short RNA-seq reads. Bioinformatics 30, 1660–1666 ([Internet] [cited 2015 Jul 2], Available from: http://bioinformatics.oxfordjournals.org/content/early/2014/02/13/bioinformatics.btu077.abstract).
Xu, Z., Wang, H., 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 35, W265–W268 ([Internet] [cited 2016 Mar 1], Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1933203&tool=pmcentrez&rendertype=abstract).
Yang, Y., Smith, S.A., 2013. Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. BMC Genomics 14, 328 ([Internet] [cited 2014 Jul 9], Available from: http://www.biomedcentral.com/1471-2164/14/328).
Zdobnov, E.M., Apweiler, R., 2001. InterProScan—an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17, 847–848 ([Internet], [cited 2016 Mar 10], Available from: http://www.ncbi.nlm.nih.gov/pubmed/11590104).
Zhang, G., Fang, X., Guo, X., Li, L., Luo, R., Xu, F., et al., 2012. The oyster genome reveals stress adaptation and complexity of shell formation. Nature 490, 49–54. http://dx.doi.org/10.1038/nature11413 (Available from: [Internet] Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved [cited 2013 Oct 31]).
Ziemert, N., Lechner, A., Wietz, M., Millán-Aguiñaga, N., Chavarria, K.L., Jensen, P.R., 2014. Diversity and evolution of secondary metabolism in the marine actinomycete genus *Salinispora*. Proc. Natl. Acad. Sci. U. S. A. 111, E1130–E1139 ([Internet], [cited 2016 Mar 18], Available from:) http://www.pnas.org/content/111/12/E1130).